doi:10.1093/plphys/kiab001 PLANT PHYSIOLOGY 2021: 0: 1–19

Plant Physiology

Research Article

Reproductive phasiRNA loci and DICER-LIKE5, but not microRNA loci, diversified in monocotyledonous plants

Parth Patel [1], Sandra M. Mathioni [2], Reza Hammond,[1] Alex E. Harkess [2], Atul Kakrana,[1] Siwaret Arikit [3], Ayush Dusia [4] and Blake C. Meyers [1,2,5,*,†]

1 Center for Bioinformatics and Computational Biology, Delaware Biotechnology Institute, University of Delaware, Newark, DE 19716, USA
2 Donald Danforth Plant Science Center, Saint Louis, MO 63132, USA
3 Department of Agronomy, Faculty of Agriculture at Kamphaeng Saen, Kasetsart University Kamphaeng Saen Campus, Nakhon Pathom 73140, Thailand
4 Department of Computer and Information Sciences, University of Delaware, Newark, DE 19716, USA
5 Division of Plant Sciences, University of Missouri, Columbia, MO 65211, USA

*Author for communication: bmeyers@danforthcenter.org (B.C.M.).
†Senior author.
These authors contributed equally (P.P., S.M.M.).
P.P. performed data analysis, generated figures, and assisted with writing the manuscript. S.M.M. gathered much of the raw sequencing data and contributed to experimental design. R.H. performed data analysis. A.E.H. performed data analysis, generated figures, and assisted with writing the manuscript. A.K. assisted with data analysis and experimental design. S.A. initiated and helped design the project, and gathered raw data. A.D. processed data. B.C.M. conceived the original research plan, contributed to the experimental design and writing of the manuscript with contributions from all authors, and supervised data analysis and figure development.
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (https://academic.oup.com/plphys) is: Blake C. Meyers (bmeyers@danforthcenter.org).

Abstract

In monocots other than maize (*Zea mays*) and rice (*Oryza sativa*), the repertoire and diversity of microRNAs (miRNAs) and the populations of phased, secondary, small interfering RNAs (phasiRNAs) are poorly characterized. To remedy this, we sequenced small RNAs (sRNA) from vegetative and dissected inflorescence tissue in 28 phylogenetically diverse monocots and from several early-diverging angiosperm lineages, as well as publicly available data from 10 additional monocot species. We annotated miRNAs, small interfering RNAs (siRNAs) and phasiRNAs across the monocot phylogeny, identifying miRNAs apparently lost or gained in the grasses relative to other monocot families, as well as a number of transfer RNA fragments misannotated as miRNAs. Using our miRNA database cleaned of these misannotations, we identified conservation at the 8th, 9th, 19th, and 3′-end positions that we hypothesize are signatures of selection for processing, targeting, or Argonaute sorting. We show that 21-nucleotide (nt) reproductive phasiRNAs are far more numerous in grass genomes than other monocots. Based on sequenced monocot genomes and transcriptomes, DICER-LIKE5, important to 24-nt phasiRNA biogenesis, likely originated via gene duplication before the diversification of the grasses. This curated database of phylogenetically diverse monocot miRNAs, siRNAs, and phasiRNAs represents a large collection of data that should facilitate continued exploration of sRNA diversification in flowering plants.

Received June 12, 2020. Accepted December 09, 2020. Advance access publication 13 January 2021

Introduction

Small RNAs (sRNAs) are key regulators of gene expression at the transcriptional and post-transcriptional level. MicroRNAs (miRNAs), a class of small noncoding RNAs with lengths ranging from 20 to 22 nucleotides (nts), are generated from stem-loop precursor RNAs processed by the RNase III family enzyme DICER-LIKE1 (DCL1), which yields a miRNA/miRNA* duplex. The duplex has 2-nt overhangs in the 3′-ends; these ends are methylated by the methyltransferase HUA ENHANCER1 for protection from degradation (Yang et al., 2006; Johnson et al., 2009). Generally, one strand of the duplex, the miRNA, is loaded into an Argonaute protein to form the RNA-induced silencing complex (RISC). RISC, via sequence homology to the miRNA, recognizes target messenger RNAs (mRNAs), and this interaction causes post-transcriptional repression by either target mRNA cleavage or translational repression (Axtell, 2013). miRNAs are involved in a multitude of plant biological processes such as seed germination (Reyes and Chua, 2007; Liu et al., 2007), leaf morphogenesis (Palatnik et al., 2003), floral development (Mallory et al., 2004), and responses to biotic (Li et al., 2011; Zhang et al., 2016) and abiotic stresses (Leung and Sharp, 2010; May et al., 2013).

Phased, secondary, small interfering RNAs (phasiRNAs), another important small RNA (sRNA) class, are distinguished from miRNAs in their biogenesis. PhasiRNA biogenesis starts from a single-stranded product of RNA Polymerase II (Pol II) derived from a genomic *PHAS* locus (a locus that makes phasiRNAs) that is capped and polyadenylated as a typical mRNA. Next, cleavage of this mRNA is directed by a 22-nt miRNA (Cuperus et al., 2010; Zhai et al., 2015). The cleaved phasiRNA precursor is made double stranded by RNA-DEPENDENT RNA POLYMERASE 6 (RDR6), and then this double-stranded RNA is processed in an iterative or "phased" manner (i.e. into consecutive 21- or 24-nt sRNAs) by a Dicer-like enzyme. DCL4 and DICER-LIKE5 (DCL5) produce 21- and 24-nt phasiRNAs, respectively (Song et al., 2012). The 21-nt phasiRNAs are widespread in plants, originating in land plants hundreds of millions of years ago as the *trans*-acting siRNA (tasiRNA) loci but also derived from diverse protein-coding gene families (Fei et al., 2013; Xia et al., 2017). The 24-nt "reproductive" phasiRNAs have been described only in angiosperms, are highly enriched in meiotic anthers, and are typically but not always triggered by miR2275 (Kakrana et al., 2018; Xia et al., 2019). A special class of 21-nt "reproductive" phasiRNAs are highly expressed in pre-meiotic anthers of some monocots, triggered by miR2118 (Johnson et al., 2009; Kakrana et al., 2018). The function of either class of phasiRNAs is still not clear, but perturbations of both result in defects in male fertility (Komiya et al., 2014; Fan et al., 2016; Ono et al., 2018; Teng et al., 2020). miRNAs have been investigated in many plant species, both in individual genomes and from limited-scale comparative analyses (Montes et al., 2014; You et al., 2017).

In the monocots, a group of about 60,000 species, most studies of sRNAs have focused on members of the Poaceae (grasses), with scant data from nongrass monocots (Kakrana et al., 2018). Rice (*Oryza sativa*), *Brachypodium distachyon* (i.e. *Brachypodium*), and maize (*Zea mays*) are the most studied of the grasses, with miRNAs characterized using varying genotypes, tissue types, growth, and stress conditions (Zhang et al., 2009; Jeong et al., 2011). With the major goal of assessing the diversity and origins of miRNAs in monocots, we analyzed sRNA data from 38 phylogenetically diverse monocots, spanning orders from the Acorales to the Zingiberales. We described sRNA size classes, miRNA conservation, divergence, sequence variability, 5′- and 3′-end nt preferences, and single-nucleotide sequence profile characterizing positional biases and providing insights into plant miRNA sequences. We performed comparative analysis of miR2118 and miR2275 and their long noncoding RNA (lncRNA) targets in monocots relative to other flowering plants, demonstrating their presence and absence in these species. We found that both miR2118 and miR2275 are conserved across diverse monocot species and are present in vegetative tissues but are found at high abundances predominantly in inflorescence tissues. The 21- and 24-nt *PHAS* loci are most numerous in the genomes of grasses, relative to other monocots, and are similarly most abundant in inflorescence tissues. Fewer *PHAS* loci were identified in nongrass monocots. Overall, our study provides a deep comparative analysis of sRNAs in monocots, including a refined database of monocot miRNAs.

Results

Sequencing from diverse monocots demonstrates atypically abundant 22-nt siRNAs

We collected materials and sequenced sRNAs from 28 monocot species spanning nine taxonomic orders: Poales (17 species), Arecales (3 species), Zingiberales (2 species), Commelinales (1 species), Asparagales (6 species), Pandanales (1 species), Liliales (1 species), Alismatales (6 species), and Acorales (1 species; Supplemental Table S1). These species included an early-diverging monocot *Acorus calamus* (Acorales) and the early diverging Poales (grasses) *Pharus parvifolius*, *Anomochloa marantoidea*, and *Streptochaeta angustifolia* (Kellogg, 2001). The sRNA libraries from these species totaled 52 vegetative and 148 inflorescence or reproductive tissues; these 200 sRNA libraries yielded 5,312,866,505 total sRNA sequences after trimming and quality control of reads (Supplemental Table S2). For some analyses, we utilized public sRNA data from an additional ten monocot species (Figure 1; Supplemental Table S1). We also included sRNAs from *Nymphaea colorata* (Nymphaeales), *Amborella trichopoda* (Amborellales), and Arabidopsis (*Arabidopsis thaliana*) as outgroups (Figure 1). In total, our study comprised billions of sRNAs from 41 diverse angiosperm species. To make these data accessible to the public, we built a series of 15 customized websites with the libraries mapped to high-quality monocot genomes, ranging from asparagus to *Zostera*; the URLs for these sites are listed in Supplemental Table 2D.

Figure 1 Phylogenetic distribution of species sampled for sRNAs. *N. colorata*, *A. thaliana*, and *A. trichopoda* were included for comparative purposes. Orders of the monocots are shaded in light blue in the legend. Red dots denote species with genome sequences available at the time of this work. The phylogeny was generated using phyloT (phylot.biobyte.de) based on NCBI taxonomy. This phylogenetic tree was annotated using iTOL (https://itol.embl.de)

We first assessed variation in the size distribution of sRNAs across the monocots. After removing reads <18 nt and >36 nt, 21- and 24-nt long sRNAs were the two dominant sizes (Figure 2; Supplemental Figure S1 and Supplemental Table S2). These 21- and 24-nt peaks typically comprise miRNAs (21 nt) and heterochromatic siRNAs (hc-siRNAs; 24 nt), but in the case of anthers may also include a large number of abundant phasiRNAs (Johnson et al., 2009; Zhai et al., 2015). The Nymphaeales and Asparagales displayed the highest relative proportions of 21- and 24-nt sRNAs compared to other taxonomic orders, respectively (Figure 2A). For the majority of sampled species, the prominent sRNA peak was 24 nt, except for the Nymphaeales and Pandanales in which the 21-nt peak was more prominent. In the Zingiberales, 21- and 24-nt size classes were similarly abundant.

To complement the global sRNA abundance analysis, we next calculated the ratio of the distinct sequences (or unique, i.e. different sequences) of 24-nt to 21-nt sRNAs; since grasses have been well characterized, we compared

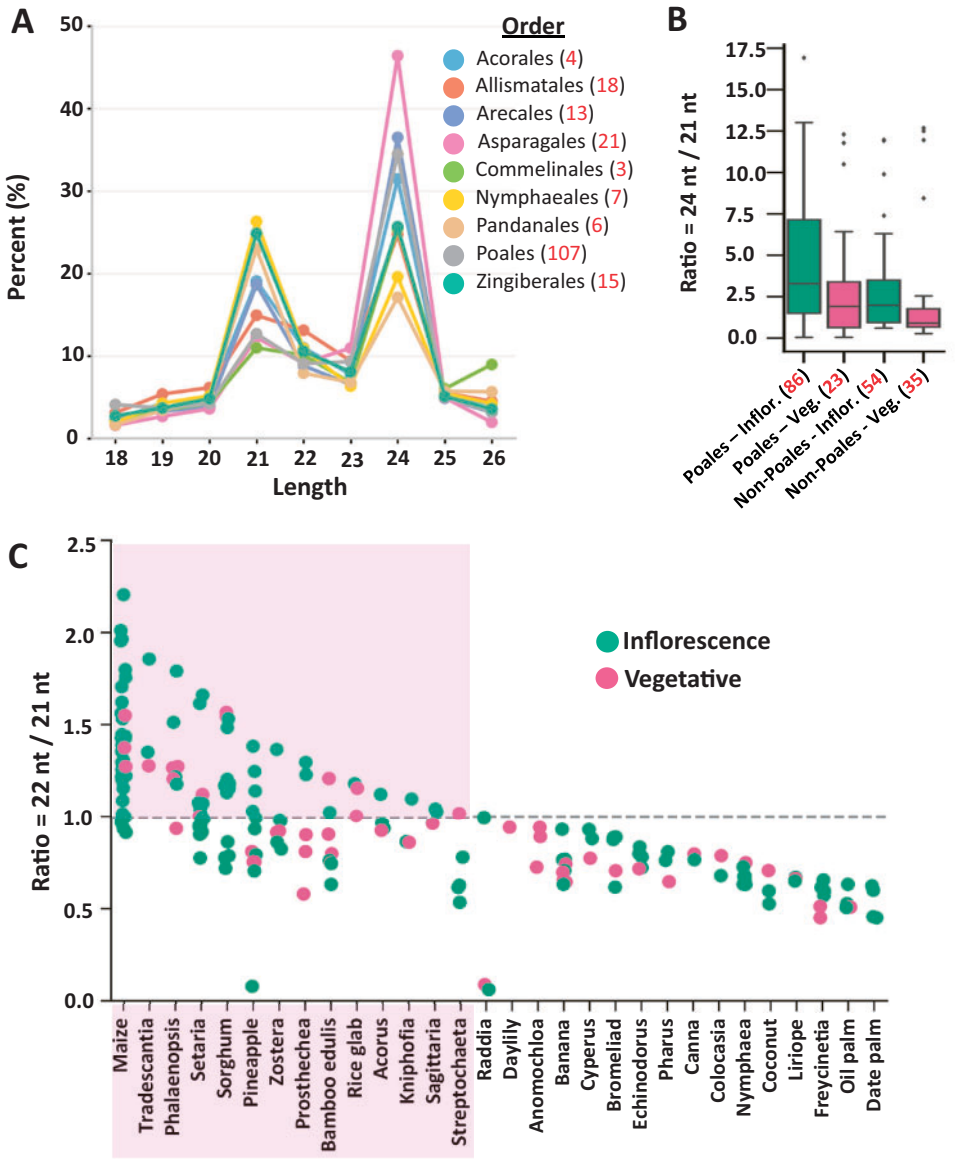

Figure 2 sRNA size distribution variation across the monocots and their tissues. A, The relative proportion of sRNAs was calculated as percentages (*Y*-axis) for each size category (*X*-axis) in total reads in 200 sequenced libraries across 28 plant species grouped in 9 different plant orders along with number of libraries (i.e. the sample size) denoted by red font. Reads longer than 26 nt were not included in this study. B, Box plots comparing the ratio of 24- and 21-nt sRNA reads (*Y*-axis) between the Poales (grasses) and non-Poales, defined as all species except the grasses (*X*-axis). The center line (black line) in each plot indicates median of the distribution. "Inflor." indicates libraries from inflorescence material, "Veg." from vegetative; red numbers indicate the number of libraries. C, Dot plot depicts the ratio of the 22- and 21-nt distinct sRNA reads (*Y*-axis) among species (*X*-axis), grouped by vegetative (pink dots) and inflorescence libraries (green dots). Highlighted pink box indicates species for which the ratio is $\geqslant 1$. Dotted gray line denotes equal ratio (value of "1" on *Y*-axis)

species in the Poales (grasses) to non-Poales (all monocot data in our study except the grasses), and inflorescence versus vegetative tissues (Figure 2B). Overall, the Poales displayed a higher proportion of 24-nt sRNAs than non-Poales across all libraries, perhaps indicative of more 24-nt hc-siRNAs or 24-nt phasiRNAs.

Next, we identified species with a disproportionately high level of 22-nt sRNAs, as our prior work has identified an unusual, RDR2-independent class of 22-nt sRNAs in maize (Nobuta et al., 2008). Our recent work using machine learning approaches demonstrated that these maize 22-nt siRNAs have distinct sequence characteristics (Patel et al., 2018). Yet, outside of these reports in maize, there has been a paucity of data on these 22-nt sRNAs in the last decade, perhaps because monocot sRNAs are so poorly characterized. We computed the ratio of the distinct 22- and 21-nt sRNAs, again comparing inflorescence and vegetative tissues (Figure 2C). This ratio exceeded 1 (i.e. higher levels of 22-nt sRNAs) for most grasses (*Setaria*, *Sorghum*, etc.) and several nongrass monocots (*Tradescantia*, *Phalaenopsis*, *Zostera*, etc.; Figure 2C). This higher proportion of 22-mers was also more often observed in inflorescence than vegetative tissues. This

result is consistent with a widespread occurrence of these poorly characterized 22-nt siRNAs in monocot species other than in just maize; their biogenesis and roles are yet to be described.

Identification of miRNAs variably conserved within the monocots

With these data, we sought to characterize miRNAs present in monocots, and those that either pre-date the split with eudicots or likely emerged since then. Since validated miRNAs longer than 22 nt are rare, our analysis focused on the 20-, 21-, and 22-nt lengths. We utilized two main strategies for the miRNA analysis. First, we identified conserved candidate miRNAs using a custom, homology-based pipeline, using mature plant miRNAs from miRBase to query all the sRNAs (see Methods section for details). This analysis yielded 84,390 candidate miRNA sequences. A spot check of these sequences in rice (for which miRNAs are well characterized) suggested that many represent low abundance variants such as sequencing errors. These sequences were filtered to find those with an abundance of $\geqslant$100 reads summed across all the libraries sequenced for a given species (i.e. low or modest accumulation in at least one sequencing library, or very low in multiple libraries). This yielded 5,354 miRNA sequences (Supplemental Table S3). These sequences belonged to 290 distinct miRNA families (all annotated in miRBase); the number of miRNA families per species is shown in Supplemental Table S3. We separated these miRNAs into those that are highly conserved, intermediately conserved, and not conserved—categories described in the following paragraphs.

Conserved miRNAs

We identified six highly conserved miRNAs found in more than 34 species; these are the well-known miRNAs miR156, miR165/166, miR167, miR171, miR319, and miR396 (Figure 3A; Supplemental Table S4, Part A). A set of another fifteen well-conserved miRNAs were found, present in 20–34 species (Figure 3A; Supplemental Table S4, Part A). Yet, another set of miRNA families was observed in 10–20 of the 41 species, which for descriptive purposes we state as having a moderate or intermediate level of conservation (Figure 3B; Supplemental Table S4, Part B). The absence of a miRNA family from one species does not imply that it is not encoded in that genome as some miRNAs are tissue-specific and our sampling and depth of sequencing was not an exhaustive analysis. However, these data were useful as a representative set of monocot miRNAs.

We made a number of observations about monocot miRNAs. First, most "three-digit miRNAs" (miRNAs with numbers lower than miR1000, i.e. miR166, miR399, miR827, etc.) are conserved across monocots (Figure 3A). Second, the data from banana showed a strong pattern for higher counts of miRNA candidates for both highly conserved (Figure 3A) and intermediately conserved (Figure 3B) groups. Third, *Colocasia*, *Echinodorus*, *Sagittaria*, and *Zostera* showed a poor representation of highly conserved miRNAs, and essentially no moderately conserved miRNAs, except for miR395 in *Colocasia* and miR1507 in *Sagittaria*. And finally, sugarcane showed weak or no representation of conserved miRNAs. The poor result in sugarcane could be attributed to the sampled tissue type, technical complication due to a highly polyploid genome, an issue of sample preparation, or reads not exceeding the abundance of 100 in the two libraries used in this analysis.

We also noted several intriguing patterns of miRNA conservation in the monocots (Figure 3B). First, miR1507, a trigger for 21-nt phasiRNAs from nucleotide binding and leucine-rich repeat pathogen-defense genes in legumes (Fei et al., 2015) is not present in the grasses (except in *Pharus*), but it is present in other monocots, perhaps indicative of a lineage-specific loss. Second, miR444, miR530, and miR1432 showed strong representation in grasses and poor representation outside of the grasses, perhaps consistent with recent evolutionary emergence. Third, miR482 was detected in early-diverged monocots but not in the grasses, consistent with earlier reports of functional diversification of miR482 and miR2118 (Xia et al., 2015). Fourth, miR894, miR8155, and miR8175 showed a similar pattern of presence across the sampled monocots, suggesting these miRNAs may comprise a family. Fifth, we identified several miRNAs present specifically as 22-mers in *Amborella*, including miR482, miR1507, and miR2275; their presence in the sister to all flowering plants suggests these miRNAs emerged prior to the monocots.

We examined several of these observations in more detail, starting with the set of three miRNA families (miR894, miR8155, and miR8175) demonstrating similar patterns of representation (Figure 3B). These miRNAs have been separately described in eudicots, mosses, and other lineages (Montes et al., 2014; Harkess et al., 2017) but not previously been shown to have a common origin. We aligned the family members and found a high degree of similarity that is suggestive of a superfamily (Supplemental Figure S2A). A review of the literature mentioning these three miRNAs found that miR894 was previously inferred to be a transfer RNA (tRNA) fragment (Montes et al., 2014). Therefore, we analyzed miR8155 and miR8175 to determine if these might also be tRNA fragments (tRFs). The annotated copy of the Arabidopsis miR8175 corresponds to the 3′-end of tRNA Asp-GTC-8-1 (Supplemental Figure S2B), while miR8155 from oil palm corresponds to the 3′-end of annotated tRNAs from multiple plants, with a BLASTN E-value of 1e–09 (e.g. alignment with *Zea mays* (chr8.trna152-MetCAT) in Supplemental Figure S2C). These misannotations may be perpetuating confusion about these sRNAs and thus the miRNAs should be blacklisted or scrubbed from databases.

We next examined the set of monocot-specific miRNAs represented by miR444, miR530, and miR1432. miR444 has been previously characterized in several grass species (Lu et al., 2008), and it was identified without further analysis in pineapple (Md Yusuf et al., 2015). We found miR444 in several other monocots, including the palms, but no earlier

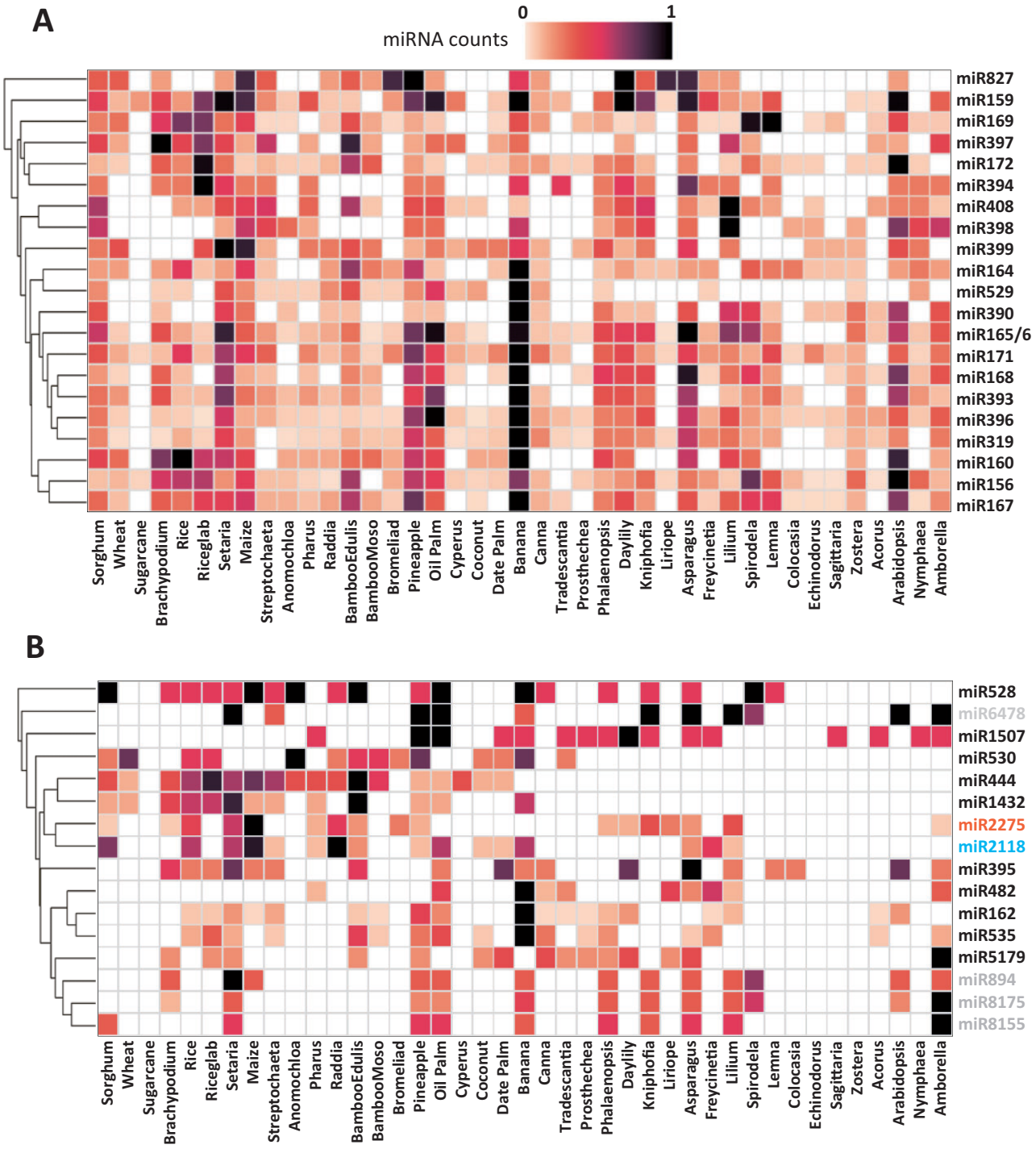

Figure 3 Variable levels of conservation in monocot-specific miRNAs. miRNA abundances were assessed using the sRNA data from vegetative and reproductive tissues. A sequence with ⩾100 reads in either vegetative or inflorescence tissues was retained. miRNA families were divided into two groups according to their conservation, defined as highly conserved and intermediate conserved. Heatmap colors represent row normalized miRNA candidates counts in the family ranging from 0 (white) to 1 (black) as illustrated in the color keys. The miRNAs were hierarchically clustered based on counts using "single" method and "Euclidean" distance. A, High conservation was defined by miRNA families identified in more than 19 species out of all 41 species examined. B, Intermediate conservation was defined by miRNA families identified between 10 and 19 species. miR2118 and miR2275, the reproductive phasiRNA triggers in the grasses, are denoted by blue and orange fonts, respectively. miR6478, miR894, miR8155, and mi8175 are in light gray as all correspond to tRFs (see main text)

diverging lineages (Figure 3B). A characteristic of miR444 in grasses is its genomic antisense configuration relative to the target gene (Lu et al., 2008); we observed the same configuration in pineapple, indicating that this arrangement may reflect its ancestral state and even its evolutionary origins (Supplemental Figure S3). miR530 and miR1432 were

also described previously in pineapple (Md Yusuf et al., 2015), and we also found that they were detected in other sister species within the commelinids, but not earlier in the monocots. Therefore, our analysis of annotated miRNAs demonstrated a combination of patterns of conservation and divergence within the monocots, with at least three monocot-specific miRNAs that emerged coincident with the commelinids.

Unannotated and evolutionarily novel miRNAs predicted as conserved within monocots

Next, we tested if our data revealed any monocot-wide, evolutionarily novel miRNAs that are not annotated in miRBase. We utilized *de novo* miRNA prediction for the monocot species for which a genome sequence is available at the time of this analysis (15 species, in 2018), and we compared these across the sRNA data of all analyzed species. This identified unannotated and weakly conserved miRNA families. For the 15 species with genomes, using sRNA data from this study and publicly available data, we used a new pipeline called *miRador* (available on Github, see Methods section) and cross-checked the results using the well-established ShortStack pipeline (Johnson et al. 2016). Both pipelines implement the strict, recently described criteria for miRNA annotation (Axtell and Meyers, 2018). In those criteria, predicted miRNAs with five or fewer nucleotide differences were then classified as members of a single miRNA family. We did not consider candidates found in only one genome, as these are harder to validate in a large-scale screening, and are thus prone to misannotation. The result of both pipelines was similar, with no candidates for novel miRNAs conserved across all 15 monocot species. There was one case for which novel miRNAs appeared to be conserved across at least two species, *Setaria* and *Sorghum* (Supplemental Table S5). These miRNAs passed our strict annotation criteria, and target prediction plus analysis of PARE data generated for this study from multiple tissues (Supplemental Table S2) found no validated targets in either genome, so their possible functions or roles in post-transcriptional silencing remain unclear. Overall, there is scant evidence for the presence of any monocot-wide, conserved and novel miRNAs.

Size distribution of conserved miRNAs displays strong conservation of length in monocots

Plant miRNAs are typically 21 or 22 nt, with this length difference often determining whether or not they function to trigger phasiRNAs; therefore, we were interested to determine whether conserved miRNAs are also conserved in their length. We computed the distribution of miRNA sizes as a percentage (between 0 and 100) of the total abundance for highly and intermediately conserved miRNA families (Figure 4). In the 37 miRNA families (Figure 4, A and B), 21 nt was the most frequent length. Four miRNA families (miR156, miR394, miR395, and miR6478) exhibited substantial proportions of 20-nt sequences (>30% of abundance). This length for both miR156 and miR394 is due to the formation of asymmetric bulges and mismatches in the duplex (Lee et al., 2015). We were curious about miR6478, predominantly a 20-nt miRNA across multiple species in our data, yet reported previously only in poplar and rice (Puzey et al., 2012; He et al., 2015). We observed that it was present in Arabidopsis and *Amborella*, but not annotated in either species. When we analyzed the sequence in the Arabidopsis genome (Supplemental Figure S4), it corresponded to tRNA precursors, and thus we infer that this is a tRF and not a true miRNA. This matches with prior conclusions about a different miRNA included in our work, miR894, also inferred to be a tRF (see above; Montes et al., 2014).

Four miRNA families (miR167, miR393, miR482, and miR2118) preferentially accumulated as 22-nt sRNAs, while others including miR1507 accumulated a large proportion of 22-nt variants (Figure 4, A and B). Length variation for miR167 and miR393 was observed previously, although miR393 was rarely reported as 22 nt in monocots, perhaps reflecting a narrow set of sampled species in this lineage (Montes et al., 2014). The 22-nt size specificity among miR482, miR1507, and miR2118 reflects well-described roles as triggers of phasiRNAs. miR2275, also a well-known trigger of phasiRNAs (Johnson et al., 2009) was, perhaps unexpectedly, not among the set of preferentially 22-nt miRNAs. The most highly abundant 21-nt sequence (zma-miR2275b-5p) turned out to be a miRNA* ("microRNA-star") sequence from these loci; this sequence has no known function or targets, so its extraordinary accumulation may reflect something unusual about miR2275, which is the only miRNA reported to date to trigger 24-nt phasiRNAs.

Single-nucleotide miRNA sequence profiles characterize position-specific nucleotide biases of conserved miRNA variants

Next, we characterized candidate miRNA sequences in greater detail at the single-nucleotide level. We first generated a nonredundant set of 2,304 candidate miRNA sequences from the set of 5,354 sequences of conserved miRNAs (in 290 distinct miRBase-annotated families) found across our libraries. We computed single-nucleotide sequence profiles for these sequences, determining the frequencies of each nucleotide (A, C, G, and U) at each position (Figure 4C). Combining these results, we made several observations: (1) in miRNA candidates, there was a 5′-nt preference for U, consistent with prior reports (Montes et al., 2014; You et al., 2017); (2) a peak of G was observed at the 8th and 9th positions; (3) in the 3′-end of the candidate miRNAs, we observed a peak of C at the 19th position (with a depletion of G and U). For comparison, we plotted the sequence profile of 3,722 distinct mature miRNAs directly from miRBase (version 21; size 20–22 nt). Several aforementioned features were conserved, except for the peak of G at the 9th and C at the 19th positions, and the miRBase miRNA sequence profile lacked distinctive sequence characteristics (Figure 4, C and D). To understand

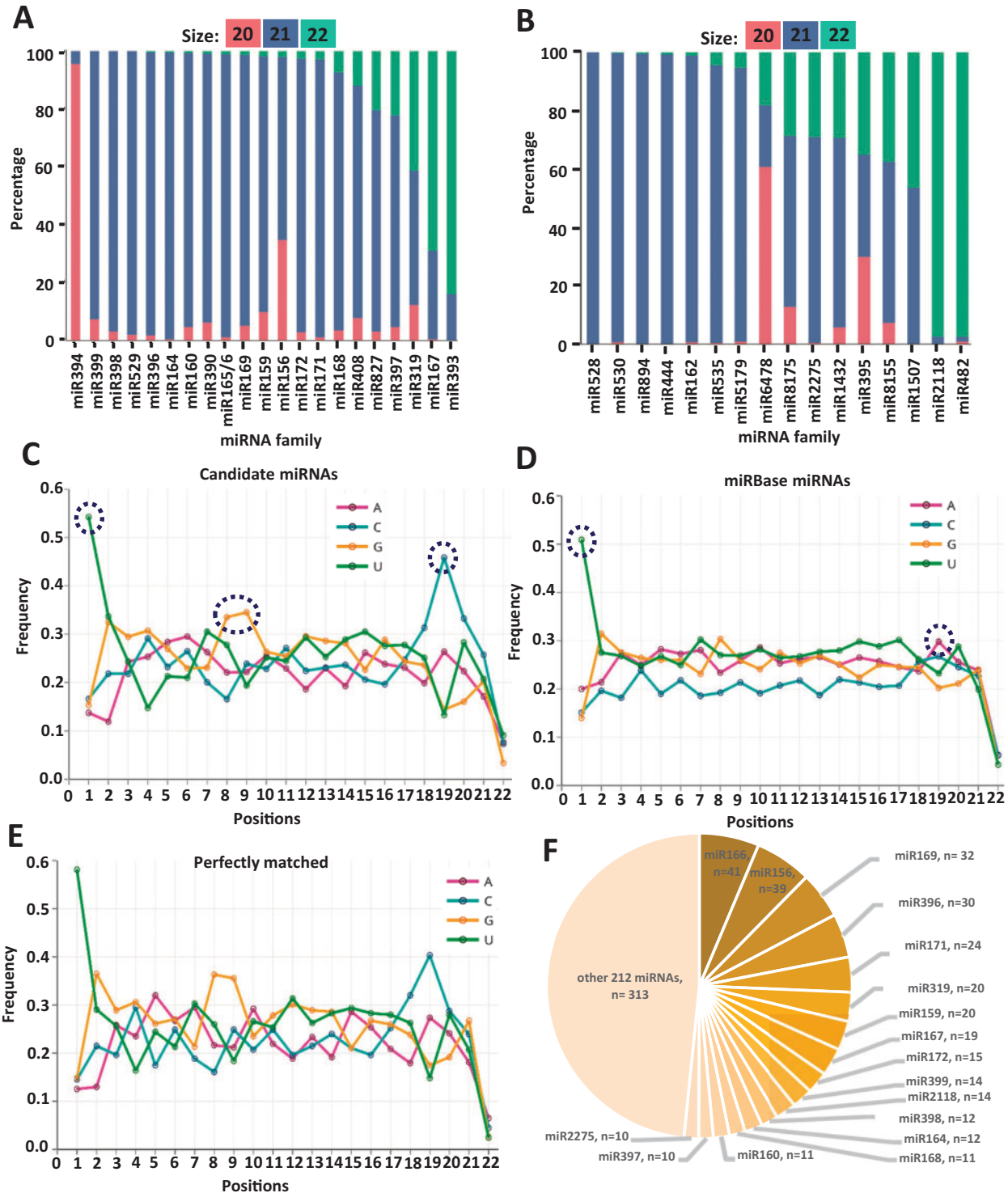

Figure 4 Size distribution and sequence profile characterize position-specific nucleotide biases of conserved miRNA variants. A, B, The stacked bar plots show the size distribution of reads as a percentage (between 0 and 100) of abundance out of total abundance (*Y*-axis, denoted 20-mers in red; 21-mers in blue; 22-mers in green) in the conserved miRNA families. As in Figure 3, conservation was defined by miRNA families identified in more than 10 species out of all 41 species examined. Bar plots are sorted from low to high percentage of 22-mers; abundances were combined for all species in which the miRNAs were detected. A, Size distribution in the most conserved 21 miRNA families (*X*-axis). B, Size distribution in the intermediate conserved 16 miRNA families. C, D, Single-nucleotide sequence profiles of unique candidate miRNAs (n =2,304) (C) and unique mature miRNA sequences (n = 3,722, size 20 to 22) from miRBase, version 21 (D). The frequencies of each of the four bases (A, C, G, and U) at each position are indicated as an open circle. E, Single-nucleotide sequence profiles of unique candidate miRNAs from (C) perfectly matching (n = 647) to the miRBase miRNAs. F, Pie chart illustrating counts of miRNA candidates 100% identical to miRBase miRNAs from (E)

the basis of this difference, we assessed the sequence profile of the subset of our 2,304 miRNAs used for Figure 4C that perfectly match miRBase-annotated miRNAs. This yielded two lists: (1) unique candidate miRNAs from panel C perfectly matching miRBase ($n = 647$, analyzed in Figure 4E) and (2) those not perfectly matching miRBase ($n = 1656$, analyzed in Supplemental Figure S5). The majority of these perfectly matched candidate miRNA sequences are well known miRNAs (Figure 4F). We observed similar sequence profiles of candidate miRNAs whether or not they perfectly matched to miRBase miRNAs (Figure 4E compared to Supplemental Figure S5); since the miRBase miRNAs lacked these distinctive signals, there may be "contaminating" annotations among miRBase miRNAs that dilute the signal. This is supported by a recent commentary (Axtell and Meyers, 2018), which suggested that miRBase in its current state contains many low confidence and erroneous annotations of miRNAs.

We next characterized and investigated the previously unreported signatures in these plant miRNAs that we observed at the 8th, 9th, and 19th positions and tested if these observations are consistent across eudicots as well. We focused on Arabidopsis miRNAs from miRBase, filtering them based on their abundance in publicly available sRNA expression data [from Gene Expression Omnibus (GEO) series GSE44622, GSE40044, GSE61362, and GSE97917]. We used a normalized abundance cutoff of 1,000 TP2M (1,000 transcripts per 2 million mapped reads) and we retained distinct sequences of size between 20 and 22, rendering a total of 138 sequences. Among these sequences, we observed the conserved pattern of G at 8th and 9th positions, but a peak of A rather than C at the 19th position (Supplemental Figure S6A). To assess the nature of this A at the 19th position ("19A"), we segregated all the miRNAs by their 5′-nt, creating four more plots. The 5′-U miRNAs, the 5′-end typical of miRNAs (Mi et al., 2008), uniquely displayed a 19C (Supplemental Figure S6B), whereas the other three classes had 19A (Supplemental Figures S6C to S6E). This 19C could be evolutionarily advantageous for 5′-U miRNAs because in a 21-nt miRNA this would yield a 5′-G on the complementary strand, i.e. the "passenger" or miRNA* strand. Since 5′-U is a strongly favored nucleotide across the majority of miRNA families and few mature miRNAs have 5′-G, this nucleotide composition may contribute to AGO sorting, loading, or binding. An alternative hypothesis is that many 5′-G/19A miRNAs may be misannotated passenger strands.

5′- and 3′-nt features of conserved miRNAs

Because the 5′-nt of miRNAs is a distinguishing feature, mainly for AGO sorting and hence function (Mi et al., 2008), we analyzed the 5′- and 3′-ends of the sRNAs described above. We characterized the 5′-nt prevalence in the 21 highly conserved and in the 16 intermediately conserved miRNA families, and focused on miRNAs from 20 to 22 nt. In the majority of miRNA families, U was the most prevalent 5′-nt, consistent with earlier reports (Montes et al., 2014; You et al., 2017; Figure 5, A and B). We observed several exceptions at the 5′-end: miR390, miR529, and miR172 predominantly displayed an A at the 5′-position, consistent with earlier reports (Montes et al., 2014; You et al., 2017). In the intermediately conserved miRNA families, a 5′-U also predominated (Figure 5B). The exceptions we found (miR6478 and miR894 with a 5′-C, and miR8155 and miR8175, with a uniform distribution of G, C, and U), were all tRFs (see above), suggesting that these 5′-nt can support the segregation of high-quality versus suspicious miRNA annotations.

The 3′-terminal nt identity of miRNAs is not well studied, although we have previously reported that this position may be more important than the seed region (positions 2–13) for miRNA-target interactions generating secondary siRNAs (Fei et al., 2015). Hence, we examined the 3′-nt prevalence in the highly and intermediately conserved miRNA families. We observed a 3′-C in the majority of miRNA families (Figure 5, C and D). A 3′-U was the second most-prevalent nucleotide (Figure 5, C and D). The 3′-nt identity in a few miRNA families was different. For example, over half of the miR398, miR399, and miR528 family members were enriched in 3′-G (Figure 5, C and D). The tRFs miR894, miR8155, and miR8175 again had terminal nucleotides inconsistent with "true" miRNAs, in this case with high proportions of 3′-A. The four 22-nt miRNA family members, known triggers of phasiRNAs (miR482, miR1507, miR2118, and miR2275) were all enriched for A in the 3′-end, a 3′-end bias perhaps required for target interactions that instigate the biogenesis of phasiRNAs (Figure 5D). Thus, 3′-nt conservation in plant miRNA families may be as important for miRNA function as the 5′-end, with special importance for triggers of phasiRNAs.

The biogenesis pathway for reproductive phasiRNAs diversified in the monocots

We next focused on a pathway that has been well described in monocots, the reproductive phasiRNAs. These phasiRNAs are enriched in anthers and require specialized miRNA triggers and lncRNA precursors, as well as the monocot-specific DCL5 protein (Zhai et al., 2015; Kakrana et al., 2018; Xia et al., 2019).

miR2118 and 21-PHAS loci

The trigger of 21-nt phasiRNAs, miR2118, was detected in 16 of the 41 species surveyed in this study (Figure 6). This miRNA is highly abundant (>100 reads) in vegetative tissues of *Raddia*, oil palm, banana, asparagus, and *Freycinetia*; in vegetative tissues of eudicots, *NLR* disease resistance genes are the most common target of the miR2118 superfamily (Zhai et al., 2011). Read abundances below 100 were observed in other monocot species, not only in the Poales but also in the Alismatales, and also in *Amborella trichopoda*. In the inflorescence tissues, miR2118 was highly abundant in most of the grasses, in the order Arecaceae, and in several other monocots (banana, asparagus, and *Freycinetia*). *Nymphaea colorata* (Nymphaea) and *Amborella* also have

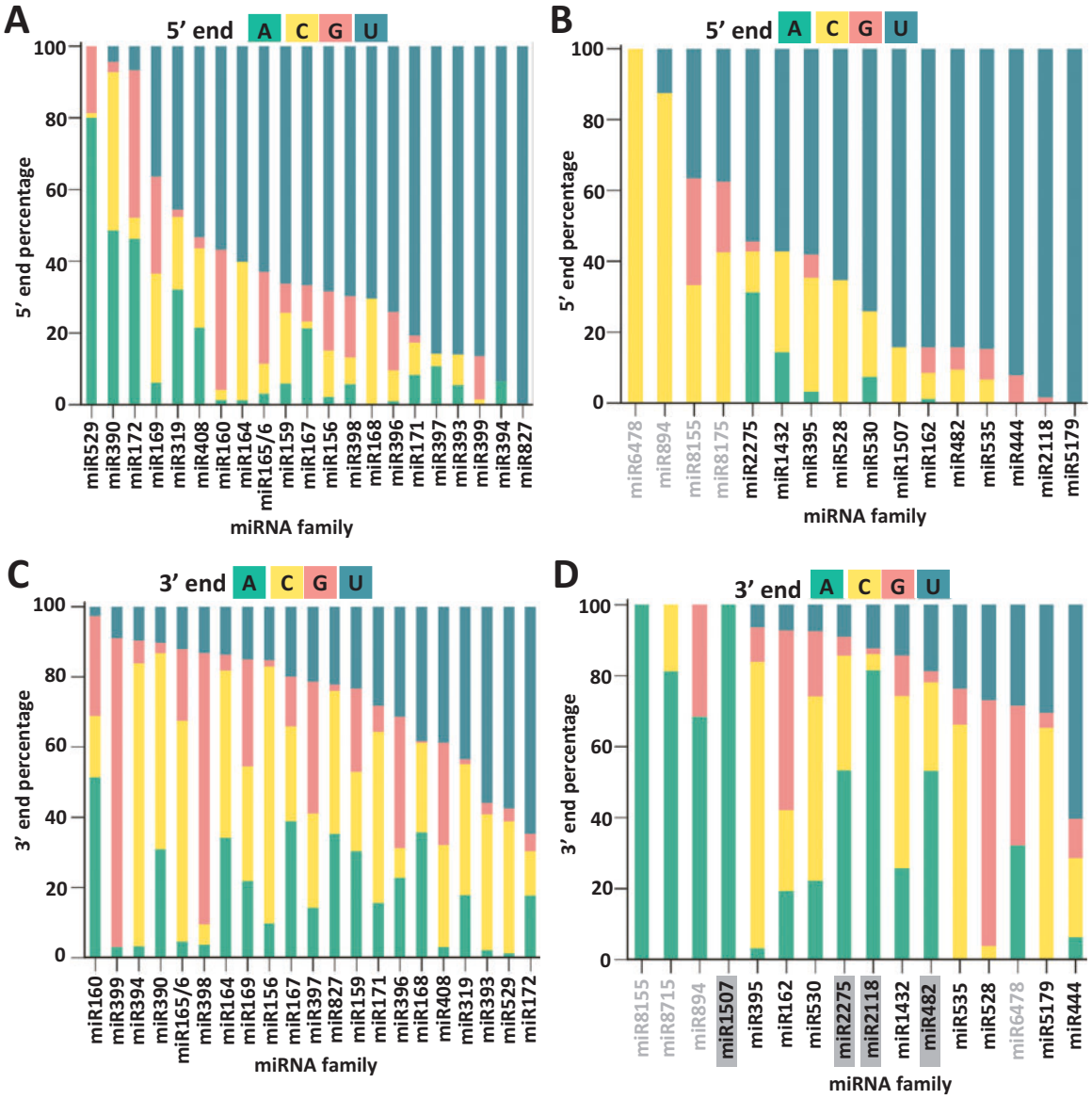

Figure 5 5′- and 3′-nt distribution of conserved miRNA families. The stacked bar plots show the 5′- and 3′-end nt composition as a percentage (between 0 and 100) of all 4 nt in the conserved miRNA families. Conservation is defined by miRNA families identified in more than 10 species out of all 41 species examined. Bar plots are sorted from low U to high U percentage. A, The 5′-nt composition (*Y*-axis) in the 21 most-conserved miRNA families (*X*-axis). B, The 5′-nt composition in the 16 intermediate-conserved miRNA families, with the misannotated tRFs noted in gray text, all enriched in 5′-C or G. C, The 3′-nt composition in the 21 most-conserved miRNA families. D, The 3′-nt composition in the 16 intermediate-conserved miRNA families. The four known triggers of phasiRNAs are highlighted with gray boxes while the tRFs are noted with gray text

miR2118 with read abundances lower than 100 (Figure 6). These results confirmed the widespread nature of miR2118 in angiosperms, with high abundance predominantly in the inflorescence tissues of monocot species outside of the grasses.

Next, we ran a phasing analysis to identify 21-nt phasiRNA-generating loci (21-*PHAS* loci) for 18 species with an available genome sequence, yielding numerous 21-*PHAS* loci (Figure 6). Grasses showed the highest counts of 21-*PHAS* loci, which were more abundant in the inflorescence tissues, with few loci showing abundance in the vegetative tissues typically, 21-nt siRNAs from *TRANS-ACTING SIRNA3* loci (i.e. *TAS3*), important for land plant development, are abundant in vegetative tissues (Xia et al., 2017). This high count of 21-*PHAS* loci in grass inflorescence tissues is consistent with their functional importance for anther fertility, for instance, as reported in rice (Fan et al., 2016). Outside of the grasses, fewer 21-*PHAS* loci were identified, and these showed similar abundances in both inflorescence and vegetative tissues, consistent with *TAS3*, as observed in pineapple, oil palm, date palm, banana, *Phalaenopsis*, asparagus, and *Amborella* (Figure 6).

We also performed *de novo* trigger prediction to characterize the miRNA triggers of 21-*PHAS* loci, to help determine

Order	Family	Species	miR2118		21-*PHAS*			
			Veg Tissue (Max Abun)	Inflo Tissue (Max Abun)	Total Count	Filtered Count	Veg Tissue	Inflo Tissue
Poales	Poaceae	*Raddia brasiliensis*	8,764	10,318				
		Phyllostachys edulis	151	1,095	446	240	4	240
		Phyllostachys heterocycla	8	NA				
		Brachypodium distachyon	0	9	419	179	4	179
		Triticum aestivum	8	0				
		Oryza glaberrima	0	0	57	7	1	7
		Oryza sativa	0	321	1,956	1,109	4	1,109
		Zea mays	0	39,912	483	272	5	272
		Sorghum bicolor	43	1,286	1,002	677	1	677
		Saccharum officinarum	0	NA				
		Setaria viridis	6	861	1,462	803	1	803
		Streptochaeta angustifolia	0	1,482	10	4	1	4
		Anomochloa marantoidea	42	NA				
		Pharus parvifolius	22	8,336				
	Bromeliaceae	*Vriesea ospinae*	0	0				
		Ananas comosus	69	49	3	1	1	1
	Cyperaceae	*Cyperus alternifolius*	0	0				
Arecales	Arecaceae	*Elaeis guineensis*	3,681	1,796	7	2	2	2
		Cocos nucifera	18	34,424				
		Phoenix dactylifera	NA	3,750	6	2	0	2
Zingiberales	Musaceae	*Musa acuminata*	1,985	7,240	13	4	4	4
	Cannaceae	*Canna indica*	0	6				
Commelinales	Commelinaceae	*Tradescantia fluminensis*	0	13				
Asparagales	Orchidaceae	*Prosthechea radiata*	0	0				
		Phalaenopsis equestris	0	17	7	2	2	2
	Xanthorrhoeaceae	*Hemerocallis lilioasphodelus*	0	0				
		Kniphofia uvaria	0	0				
	Asparagaceae	*Liriope muscari*	0	0				
		Asparagus officinalis	345	1,188	31	19	19	19
Pandanales	Pandanaceae	*Freycinetia cumingiana*	1,023	1,062				
Liliales	Liliaceae	*Lilium maculatum*	0	217				
Alismatales	Araceae	*Spirodela polyrhiza*	0	NA	0	0	0	0
		Lemna gibba	0	NA	0	0	0	0
		Colocasia esculenta	33	0				
	Alismataceae	*Echinodorus uruguayensis*	0	0				
		Sagittaria montevidensis	8	0				
	Zosteraceae	*Zostera marina*	0	10	0	0	0	0
Acorales	Acoraceae	*Acorus calamus*	0	0				
Brassicales	Brassicaceae	*Arabidopsis thaliana*	0	0				
Nymphaeales	Nymphaeaceae	*Nymphaea colorata*	0	14				
Amborellales	Amborellaceae	*Amborella trichopoda*	25	33	10	5	5	5

Figure 6 miR2118 abundance and 21-*PHAS* counts in monocots. The miR2118 maximum abundance is shown for each monocot tissue and species where it was identified. The total and filtered 21-*PHAS* counts are shown for each monocot tissue and species where they were identified. The filtered 21-*PHAS* counts for vegetative (Veg) and inflorescence (Inflo) tissues are also shown. Color legend for miR2118 abundance: light blue <100 reads, dark blue $\geqslant 100$ reads. NA = not available

their potential roles and biogenesis. In the grasses, the biogenesis of reproductive phasiRNAs was typically dependent on one of two 22-nt miRNA triggers (miR2118 and miR2275, for 21- and 24-nt phasiRNAs), while *TAS3* is triggered by miR390. The majority of the 21-*PHAS* loci in the grasses were triggered by miR2118 (Table 1). In the monocots outside of grasses, there were comparatively few 21-*PHAS* loci, and only a subset of these had miR2118 as trigger, indicating other miRNAs may function as triggers; this is not unexpected as there are many miRNAs that trigger 21-

Table 1 The 21- and 24-*PHAS* loci with miRNA triggers for sampled genera

Order	Family	Species	Total count of 21-*PHAS* loci	No. of loci with miR2118 as a trigger	Total count of 24-*PHAS* loci	No. of loci with miR2275 as a trigger
Poales	Poaceae	*Phyllostachys edulis*	446 (4)	373	25	16
		Brachypodium distachyon	419 (2)	362	217	186
		Oryza sativa	1956 (9)	1810	126	101
		Zea mays	483 (6)	395	246	134
		Sorghum bicolor	1002 (1)	921	256	93
		Setaria viridis	1462 (4)	1351	383	205
		Streptochaeta angustifolia	10 (1)	3	1	0
	Bromeliaceae	*Ananas comosus*	3 (1)	2	89	0
Arecales	Arecaceae	*Elaeis guineensis*	7 (2)	1	4	0
		Phoenix dactylifera	6 (0)	0	21	0
Zingiberales	Musaceae	*Musa acuminata*	13 (1)	3	6	1
Asparagales	Orchidaceae	*Phalaenopsis equestris*	7 (1)	0	0	0
	Asparagaceae	*Asparagus officinalis*	31 (1)	10	54	2
Allismatales	Araceae	*Spirodela polyrhiza*	0	0	0	0
		Lemna gibba	0	0	0	0
	Zosteraceae	*Zostera marina*	0	0	21	0
Amborellales	Amborellaceae	*Amborella trichopoda*	10 (1)	2	1	0

Species are listed only for which a genome was available, and a single representative for each genus is shown. The number of 21-*PHAS* loci that are *TAS3* loci is denoted in parenthesis

nt phasiRNAs from diverse targets (Fei et al., 2013). In fact, a proportion of these were *TAS3* (Table 1). However, the relative absence of 21-nt reproductive phasiRNAs outside of the grasses suggests that they are less prevalent, although it is possible that we did not sample the correct anther stage.

Finally, since the number of 21-*PHAS* loci is quite high in the grasses, we asked whether these loci might be important to have on all chromosomes, for some undescribed role in chromosome biology (e.g. chromosome pairing). We analyzed the chromosomal distribution of 21-*PHAS* loci for five grasses which we had high quality genomes (*Brachypodium*, rice, *Sorghum*, maize, and *Setaria*). This showed an enrichment for 21-*PHAS* loci in some chromosomes, and a presence of at least one locus on all chromosomes (Supplemental Figure S7). While this is consistent with a hypothesis in which these loci are important for some sort of *cis* activity within the chromosome, no association was observed between the size of the chromosome and the count of its 21-*PHAS* loci (Supplemental Figure S7). However, this hypothetical activity would likely be limited to the grasses, given that nongrass monocots had few loci.

miR2275 and 24-PHAS loci

The only known function of miR2275 is to trigger biogenesis of 24-nt reproductive phasiRNAs. We detected this miRNA in *Amborella* and found it in both vegetative and inflorescence tissues (Figure 7). *Acorus*, a basal monocot, and *Nymphaea*, a basal angiosperm, had low abundances of miR2275 in the inflorescence tissues. The early diverging grasses, *Pharus* and *Streptochaeta*, also showed moderate abundances of miR2275 in the inflorescence tissues. The highest abundances of miR2275 were in grass inflorescences (Figure 7), while in vegetative tissues the abundance was highest in *Amborella*, *Phalaenopsis*, and bamboo. The high abundance of miR2275 in grasses may reflect an increased utilization of the pathway generating 24-nt reproductive phasiRNAs.

We next identified 24-*PHAS* loci for 15 species for which a genome sequence was available. The highest count of 24-*PHAS* loci was in the grasses (Figure 7), although only a single 24-*PHAS* locus was identified for the early-diverged grass *Streptochaeta*. Outside the grasses, pineapple and asparagus had the highest counts of 24-*PHAS* loci, with numbers of loci that were similar to maize. We found no homologs of miR2275 in the sea grass *Zostera* or date palm, although we identified 24-*PHAS* loci in both species (12 and 7 loci, respectively), perhaps consistent with asparagus or Solanaceous species (see below). We analyzed the chromosomal distribution of 24-*PHAS* loci in five grasses (*Brachypodium*, rice, *Sorghum*, maize, and *Setaria*), and as with 21-*PHAS* loci, 24-*PHAS* loci were found on all chromosomes, with counts elevated on some chromosomes (Supplemental Figure S7). Again, we observed no association between the size of the chromosome and the number of 24-*PHAS* loci.

We then predicted the triggers of 24-*PHAS* loci (Table 1). miR2275 was the trigger for the majority of 24-*PHAS* loci in the grasses. miR2275 was also generally not the trigger for 24-*PHAS* loci outside grasses. For example, in *Zostera*, 21 of the 24-*PHAS* loci showed no evidence that miR2275 was their trigger (Figure 7). This is consistent with recent

Order	Family	Species	miR2275		24-*PHAS*			
			Veg Tissue (Max Abun)	Inflo Tissue (Max Abun)	Total Count	Filtered Count	Veg Tissue	Inflo Tissue
Poales	Poaceae	*Raddia brasiliensis*	10	14,196				
		Phyllostachys edulis	1,031	127	25	15	0	15
		Phyllostachys heterocycla	69	NA				
		Brachypodium distachyon	9	107	217	145	0	145
		Triticum aestivum	0	0				
		Oryza glaberrima	16	58	4	1	0	1
		Oryza sativa	0	1,428	126	77	0	77
		Zea mays	0	91,833	246	160	0	160
		Sorghum bicolor	0	2,080	256	162	0	162
		Saccharum officinarum	0	NA				
		Setaria viridis	7	8,956	383	273	0	273
		Streptochaeta angustifolia	0	86	1	1	0	1
		Anomochloa marantoidea	0	NA				
		Pharus parvifolius	0	243				
	Bromeliaceae	*Vriesea ospinae*	0	2,482				
		Ananas comosus	0	138	89	42	0	42
	Cyperaceae	*Cyperus alternifolius*	0	11				
Arecales	Arecaceae	*Elaeis guineensis*	0	38	4	1	0	1
		Cocos nucifera	0	25				
		Phoenix dactylifera	NA	0	21	7	0	7
Zingiberales	Musaceae	*Musa acuminata*	26	28	6	3	0	3
	Cannaceae	*Canna indica*	0	29				
Commelinales	Commelinaceae	*Tradescantia fluminensis*	0	14				
Asparagales	Orchidaceae	*Prosthechea radiata*	0	9				
		Phalaenopsis equestris	145	13	0	0	0	0
	Xanthorrhoeaceae	*Hemerocallis lilioasphodelus*	0	368				
		Kniphofia uvaria	16	698				
	Asparagaceae	*Liriope muscari*	10	666				
		Asparagus officinalis	9	956	54	32	0	32
Pandanales	Pandanaceae	*Freycinetia cumingiana*	0	0				
Liliales	Liliaceae	*Lilium maculatum*	0	661				
Allismatales	Araceae	*Spirodela polyrhiza*	0	NA	0	0	0	0
		Lemna gibba	0	NA	0	0	0	0
		Colocasia esculenta	0	0				
	Allismataceae	*Echinodorus uruguayensis*	0	24				
		Sagittaria montevidensis	0	0				
	Zosteraceae	*Zostera marina*	0	0	21	12	0	12
Acorales	Acoraceae	*Acorus calamus*	0	57				
Brassicales	Brassicaceae	*Arabidopsis thaliana*	0	0				
Nymphaeales	Nymphaeaceae	*Nymphaea colorata*	0	50				
Amborellales	Amborellaceae	*Amborella trichopoda*	320	135	1	0	0	0

Figure 7 miR2275 abundance and 24-*PHAS* counts in monocots. The miR2275 maximum abundance is shown for each monocot tissue and species in which it was identified. The total and filtered 24-PHAS counts are shown for each monocot tissue and species in which they were identified. The filtered 24-*PHAS* counts for vegetative (Veg) and inflorescence (Inflo) tissues are also shown. Color legend for miR2275 abundance: beige <100 reads, orange $\geqslant 100$ reads. NA = not available

observations from work in garden asparagus (Kakrana et al., 2018) and tomato (Xia et al., 2019), both of which have 24-nt reproductive *PHAS* loci that lack obvious miRNA triggers and indicate a diversity of initiation mechanisms for 24-*PHAS* loci outside of the grasses. In garden asparagus, 24-nt phasiRNAs may be derived from inverted repeats even without miR2275, although how this initiates is unclear (Kakrana et al., 2018). A complete understanding of the diversity of biogenesis mechanisms for 24-*PHAS* loci will require detailed studies in these species.

DCL5 divergence in the monocots

Earlier work on 24-nt reproductive phasiRNAs has identified a key role for DCL5 in their biogenesis (Teng et al., 2020). This Dicer-like protein has not been well characterized outside of the grasses (or even within the grasses), so we sought to take advantage of sequenced monocot genomes and transcriptomes to determine when DCL5 emerged, and if it might coincide with any patterns of reproductive phasiRNA expression. We identified putative orthologs of DCL3 and DCL5 encoded in 12 monocot genomes plus *Amborella* and *Nymphaea*, extracted the predicted protein sequences, and performed a phylogenetic analysis (Figure 8). DCL5 orthologs were only present in *Dioscorea* and more recently diverged monocot lineages. We hypothesize that DCL5 evolved via a DCL3 tandem duplication event or whole genome duplication event, and that DCL5 emerged sometime before the diversification of *Dioscorea*. Intriguingly, DCL5 appeared to have been lost independently in some orders, such as the Asparagales (*Asparagus officinalis* and *Dendrobium officinale*). Dicer-like sequences are typically long (~1,600 amino acids) with many small exons, and thus are prone to misannotation. Manual searches of the *A. officinalis* genome sequence (Harkess et al., 2017), instead of relying solely on published annotations, revealed a partial Dicer-like sequence with homology to DCL5 (Supplemental Figure S8). However, the Asparagales still express 24-nt reproductive phasiRNAs triggered by miR2275 (Table 1), suggesting that DCL3 may have retained a redundant ancestral function. One explanation may be that species in the Asparagales have lost DCL5 but have evolved modified genomic substrates for DCL3 and miR2275 to still produce 24-nt reproductive phasiRNAs. In *A. officinalis*, 24-nt reproductive phasiRNAs are often derived from inverted repeats, which may have evolved as a DCL5-independent mechanism to produce phasiRNAs (Kakrana et al., 2018). We conclude that monocots which emerged coincident with the duplication of DCL3 may have adapted diverse mechanisms for production of 24-nt phasiRNAs. Earlier diverged species including eudicots (Xia et al., 2019) likely utilized DCL3, while the specialized DCL5 emerged in later diverged species, including the grasses.

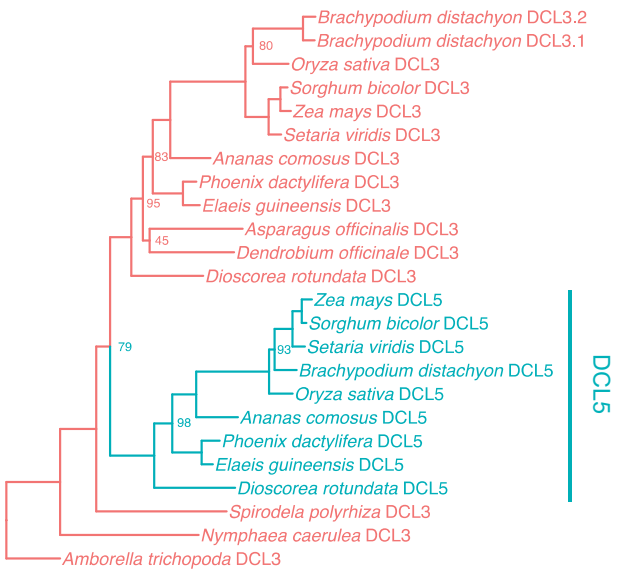

Figure 8 The origin of DCL5 in the monocots. A maximum likelihood (RAxML tree of monocot-wide DCL3 and DCL5 protein sequences. Only bootstrap values less than 100 are presented on nodes

Discussion

Our understanding of plant sRNAs is largely derived from work focused on species with sequenced genomes, including Arabidopsis, rice, soybean, Medicago, etc. A small number of studies have surveyed more diverse species, often lacking genomic data, such as lycophytes, ferns, and diverse angiosperms (Montes et al., 2014; You et al., 2017). We focused on a poorly sampled but diverse group of angiosperms, the monocots. We analyzed 41 species, including 38 monocot species, ranging from *Acorales* and *Arecales* to *Zingiberales*, totaling 308 sRNA libraries, including 200 sRNA libraries that were newly generated. Some observations were not surprising: the predominant size classes, 21 and 24 nt, are typical of plants. Yet, by calculating the ratio of 24- to 21-nt sRNAs, we found a higher ratio in the grasses and in the inflorescence tissues compared to nongrasses and vegetative tissues, respectively. We observed in the inflorescence libraries a disproportionately high level of 22-nt sRNAs compared to vegetative tissues for most grasses (*Setaria*, *Sorghum*, etc.) and for several nongrass monocots (*Tradescantia*, *Phalaenopsis*, *Zostera*, among others). There were both similarities and differences with maize; we demonstrated the significant presence of 22-nt siRNAs outside of maize, but also found that these 22-nt siRNAs have a distinct sequence composition relative to maize (Patel et al., 2018). The biogenesis and functions of these 22-nt siRNAs remains unclear. Perhaps in these species, DCL2 has a role in silencing endogenous elements, as it is the primary Dicer-like protein that produces 22-nt sRNAs, at least in Arabidopsis (Blevins et al., 2006).

Based on sequence homology, we characterized 37 miRNA families (21 highly and 16 intermediately conserved) and observed miRNA conservation patterns such as lineage-specific loss (e.g. miR1507), recent evolutionary emergence (e.g. miR444, miR530, and miR1432), functional diversification (absence of miR482 and emergence of miR2118 in the grasses), and emergence prior to monocots (presence of miR482, miR1432, and miR2275 in *Amborella*). Our characterization of conserved miRNAs using single-nucleotide miRNA sequence profiles revealed a position-specific nucleotide biases (at the 8th, 9th, and 19th positions) of conserved miRNA variants, potentially influencing AGO sorting. Since these analyses of bias are exquisitely sensitive due to the numerous sequences that were analyzed, they may reflect the influence of selection, perhaps for interactions with AGO proteins, making functional analysis difficult. Lastly, we did not identify any significant pattern of novel, previously unannotated miRNAs that are conserved across all monocots.

Our analysis of miRNA sequence characteristics identified a number of issues with miRBase-annotated miRNAs. The analysis of sequence profiles among conserved miRNAs identified numerous strong indicators of true miRNAs. This includes the well-described 5′- and 3′- nt, but there are internal nucleotides that may be important, with previously unrecognized characteristics such as the peak of G that we observed at the 8th and 9th positions, and the peak of C at the 19th position (with a depletion of G and U). We also confirmed prior observations that misannotations persist in miRBase and may cause recurring annotation problems, such as the set of tRFs represented by miR6478 and miR894/miR8155/miR8175. There is perhaps a need to track not just true annotations of miRNAs but also to track false annotations along with the explanation of why these sequences are deprecated. Our work supports the case for making community-driven improvements to miRBase (Axtell and Meyers, 2018). A related improvement would be an automated interface to provide rapid quality assessment of miRNAs, assign unique miRNA identifiers, and track targets including whether or not the loci yield phasiRNAs.

Finally, we showed that the miR2118 and miR2275 triggers of reproductive phasiRNAs, and their associated genomic *PHAS* loci, are prevalent in angiosperms. This overall observation is consistent with recent work on eudicots (Xia et al., 2019). Moreover, we demonstrated conservation of miR2118 and miR2275 across monocot species, with evidence of expression in the inflorescence tissues of all grasses, as well as limited expression in some vegetative tissues in a few non-grasses. Similarly, we concluded that 21- and 24-*PHAS* loci (the sources of reproductive phasiRNAs) are particularly numerous and abundantly expressed in inflorescence tissues of the grasses. *In silico* trigger identification in the grasses determined that miR2118, miR390, and miR2275 are triggers of most 21-*PHAS* loci, the handful of *TAS3* loci, and 24-*PHAS* loci, respectively. Prior work in species outside of the grasses has also demonstrated that 24-nt reproductive phasiRNAs may be generated by noncanonical pathways (Kakrana et al., 2018; Xia et al., 2019). In fact, this variation in the utilization of miR2275 may reflect an evolutionary period of divergence in 24-nt reproductive phasiRNA biogenesis, coincident with changes occurring in the divergence of DCL3 and DCL5. We narrowed the phylogenetic placement in monocot evolution during which DCL5 likely emerged. In the lineages that diversified following the evolution of DCL5, there was presence/absence variation for DCL5, suggesting that neofunctionalization was incomplete and its necessary function in reproductive phasiRNA biogenesis observed in grasses (Song et al., 2012) was not yet fixed. Future functional studies should test the specialization and activity of DCL5 from the genomes of these first lineages to inherit it.

In conclusion, we are in a period of rapid, large-scale data acquisition, including both genomes and the transcript data from these genomes. The interpretation of these data requires increasingly sophisticated methods amenable to high-throughput analyses. One particularly intriguing branch of research focuses on machine learning applications, which in our experience is transforming sRNA informatics (Patel et al., 2018). Future applications of machine learning may provide even deeper insights into comparative analysis, elucidating uncharacterized aspects of sRNA sequences, and elevating our understanding of miRNAs, phasiRNAs, and other classes of plant RNAs.

Methods

Plant material

Plant materials were collected from various locations and are detailed in Supplemental Table S1. Plant tissues were dissected manually and, when necessary, using a stereomicroscope for magnification and a 2-mm stage micrometer (Wards Science, cat. #949910). Tissues were flash frozen in liquid nitrogen and stored at −80°C until RNA extraction was performed.

RNA extraction

Samples were ground in cold mortars and pestles using liquid nitrogen. Total RNA isolation was performed using the Plant RNA Reagent (Thermo Fisher, cat. no. 12322012). For anther and other tissues with limited amount of plant material, we used the detailed protocol for RNA extraction previously published (Mathioni et al., 2017).

sRNA size selection, library preparation, and sequencing

sRNA size selection and library preparation using the TruSeq Small RNA Library Prep Kit (RS-200-0012, Illumina) were performed following the detailed protocol previously published (Mathioni et al., 2017). All sRNA libraries were sequenced using single-end mode with 51-nt reads, on an Illumina HiSeq 2500 Instrument in the University of Delaware Sequencing and Genotyping Center at the Delaware Biotechnology Institute.

PARE library preparation and sequencing

PARE libraries were constructed as described previously (Zhai et al., 2014), with modifications as follows: the amount of total RNA used as starting material was 10–20 μg, depending on the availability of each sample; the incubation time for the 5′-adapter ligation was 2 h; the incubation time for the reverse transcription was 2 h; the second strand cDNA synthesis was performed with 10 cycles instead of 7 cycles; the incubation time for the MmeI digestion was 2 h; the incubation time of the 3′-double-strand DNA adapter ligation was performed overnight (∼8 h); the final PCR amplification of the PARE library was performed with 18 cycles instead of 15 cycles. These changes were necessary because the total RNA starting amount was much lower than the amount recommended in the original protocol (40–75 μg). All the other parameters were kept as described in the original protocol. All PARE libraries were sequenced using single-end mode with a 51-nt read length on an Illumina HiSeq 2500 Instrument in the University of Delaware Sequencing

and Genotyping Center at the Delaware Biotechnology Institute.

sRNA and PARE data processing

The sRNA data were processed as previously described (Mathioni et al., 2017). The PARE data were processed using sPARTA as previously described (Kakrana et al., 2014) with updates available at https://github.com/atulkakrana.

MicroRNA analysis

We used two strategies to analyze the miRNAs. First, we generated a unique list of miRBase entries from version 21, with all mature miRNA sequences from the available plant species. We used this list to query our sRNA dataset for mature miRNA sequences. The criteria used for the query were based on sequence similarity (see below). In the second strategy, we ran a *de novo* miRNA prediction using both a new prediction package called *miRador* (https://github.com/rkweku/miRador) and using ShortStack (Johnson et al., 2016) for the each of species with genome sequences available.

Homology-based identification of miRNA sequences in diverse monocots using BLAST

Due to a lack of sequenced genomes in monocots, we conducted a homology-based search to identify miRNAs in 40 monocot species. We added Arabidopsis (*A. thaliana*) as a control, making it a list of 41 species. We merged all sRNA libraries from these 41 species into one and built a BLAST database (BLASTdb; Altschul et al., 1990). The command line used was makeblastdb -in file -out name -dbtype nucl -title title. Each sequence name in this file embeds information about the species name, raw read count, and tissue (vegetative or inflorescence). To compare our sRNA sequences in the BLASTdb, we collapsed a Viridiplantae-specific, nonredundant, mature miRNA sequences from miRBase version 21 (Kozomara and Griffiths-Jones, 2014). This set of unique miRNA sequences completed our reference set. To identify miRNAs present in our BLASTdb, we aligned the BLASTdb to the reference set using BLASTN. The command line used was blastn -query file -strand plus -task blastn-short -db name -out file -perc_identity 75 -num_alignments 200 -no_greedy -ungapped -outfmt "6 qseqid query sseqid pident length qlen qstart qend slen sstart send gaps mismatch positive evalue bitscore". BLASTN was performed with –ungapped and -no_greedy to facilitate end-to-end and nongreedy alignment, respectively. Here, we utilized "blastn-short", which is optimized for sequences less than 30 nt. We used 75% percent identity for BLASTN sequence scan to account for $\leqslant 4$ mismatches between a mature miRNA from the reference set and its homolog (subject sRNA sequence) in the BLASTdb. Output file contained fields identified as qseqid, query, sseqid, pident, length, qlen, qstart, qend, slen, sstart, send, gaps, mismatch, positive evalue, and bitscore in a TAB separated format.

To process miRNA annotation results from BLAST, the output from BLAST was filtered to determine the valid homologs. We used a custom python script for this filtration process. The filtration process was as follows: (1) we processed the tabular results from BLAST to compute 5′- and 3′-end overhangs, mismatches, matches, and total variance. Total variance was a sum of the nucleotides that were not aligned, including no >2 nt on 5′- and 3′-end overhangs and mismatches ($\leqslant 4$). The total variance cutoff was set to 5. Subject sequences (aka candidate homologs) satisfying this cutoff were given a status of "pass", otherwise "fail". In addition to the output from BLAST, we added extra columns (hang5, hang3, match, mismatch, unalign, totalvariance, and status). (2) We kept only the subject sRNA sequences that were of length between 20 and 22 nt. (3) When a subject sRNA sequence matched two or more mature miRNAs from the reference set, the best match was determined as the alignment that contained the highest bitscore. (4) For each sRNA subject sequence that passed these criteria, we determined the raw read count in the inflorescence and in the vegetative tissues across all sRNA libraries used. All of the potential homolog candidates were chosen based on their raw read counts more than 99 reads in either vegetative or inflorescence tissues. We used another custom python pipeline to obtain these read counts and added three extra columns (Homologous Sequence, Vegetative Raw Read Count, and Reproductive Raw Read Count) to yield the final results.

To obtain the count of miRNA families, we used three-letter miRBase (v21) codes (i.e. identifiers lower than miR1000) and generated the set of Viridiplantae-specific miRNA families using a custom python script. The same script searched the list of entries from the final results and collapsed these candidate sequences using a column query into the list of miRNA families, demonstrating the conservation of families of miRNAs in these monocot species.

De novo miRNA prediction for novel miRNAs

sRNA libraries were trimmed (Patel et al., 2015) and mapped to their respective genomes using Bowtie (Langmead et al., 2009). miRNAs were then predicted in all libraries utilizing a version of miREAP software (https://sourceforge.net/projects/mireap/). These miRNAs were assessed for their similarity to known miRNAs using BLAST. Predicted miRNAs that had five or fewer differences were then classified as members of those miRNA families.

PhasiRNA analysis

PhasiRNA (*PHAS*)-generating loci were identified using the *PHASIS* pipeline (https://github.com/atulkakrana/PHASIS; Kakrana et al., 2017). Triggers for these *PHAS* loci were further identified using the *phastrings*, a component of the *PHASIS* pipeline.

Dicer-like gene family and phylogeny

De novo gene families were circumscribed using a set of diverse monocot genomes from Phytozome v12.1 (*Ananas comosus, A. officinalis, Brachypodium distachyon, Musa acuminata, Oryza sativa, Sorghum bicolor, Zostera marina,*

Setaria viridis, and *Zea mays*), plus *Amborella trichopoda)* using OrthoFinder (v2.2.1; Emms and Kelly, 2015). DCL3 and DCL5 proteins were contained in a single orthogroup. Additional monocot sequences were added from manual BLASTP searches of the *Dioscorea rotundata* genome (Tamiru et al., 2017), and from NCBI for *Elaeis guineensis*, *Phoenix dactylifera*, and *Dendrobium catenatum*.

Only proteins that were full, canonical Dicer-like proteins (including a helicase domain, Dicer domain, PAZ, and two tandem RNAse III domains) were retained, except in a single case of possibly misannotated recent DCL3 paralogs in *Brachypodium*. This led to the exclusion of several Dicer-like genes in the analysis, which may be pseudogenes, incorrectly assembled or unannotated genomic regions, or noncanonical Dicer-like proteins that perform a currently unknown function (Supplemental Figure S8).

Complete proteins were aligned using default settings in PASTA (v1.6.4; Mirarab et al., 2015), followed by a maximum likelihood gene tree using RAxML (v8.2.11; Stamatakis, 2014) over 100 rapid bootstraps with options "-x 12345 –f a –p 13423 –m PROTGAMMAAUTO". Trees were visualized and manipulated in ggtree (Yu et al., 2017).

Accession numbers

The accession numbers from Genbank, mainly the GEO, for all the sRNA libraries generated in this study and all the libraries from published studies and used in this study for comparison purposes are listed in Supplemental Table S2. The URLs for websites for direct public access to these libraries, mapped onto reference genomes, are included in this table.

Supplemental data

The following materials are available in the online version of this article.

Supplemental Figure S1. Size distribution of sRNA sequences in nine different plant orders used in this study.

Supplemental Figure S2. Sequence analysis of homologs miR894, miR8155, and miR8175 identifies tRNA homology.

Supplemental Figure S3. The genomic structure of the natural antisense microRNA miR444 is conserved in pineapple.

Supplemental Figure S4. miR6478 is derived from a tRNA precursor in Arabidopsis.

Supplemental Figure S5. Single-nucleotide sequence profiles of unique candidate miRNAs.

Supplemental Figure S6. Arabidopsis 5′ U miRNAs share a distinctive C in the 19th position with high-confidence monocot miRNAs.

Supplemental Figure S7. Count of 21-and 24-*PHAS* loci per chromosome in five grass species.

Supplemental Figure S8. DCL3- and DCL5-like sequences and predicted domains from across the monocots based on genome annotations.

Supplemental Table S1. Detailed information for the 41 species used in this study.

Supplemental Table S2. Data used in this study.

Supplemental Table S3. Number of miRNA families per species.

Supplemental Table S4. Highly and moderately conserved miRNA families.

Supplemental Table S5. miRNA candidates conserved across monocot species.

Acknowledgments

We would like to thank Robert J. Orth and Andrew J. Johnson for providing *Zostera marina* samples; Elizabeth Kellogg for providing *Streptochaeta angustifolia* and *Pharus parvifolius* samples; Malia Gehan for providing *Setaria viridis* seeds; Nadia Shakoor for providing *Sorghum bicolor* seeds; the Longwood Gardens (https://longwoodgardens.org/) for providing samples from *Vriesea ospinae*, *Cyperus alternifolius*, *Canna indica*, *Tradescantia fluminensis*, *Prosthechea radiata*, *Freycinetia cumingiana*, *Echinodorus uruguayensis*, *Sagittaria montevidensis*, and *Nymphaea colorata*; the Chanticleer Garden (http://www.chanticleergarden.org/) for providing *Acorus calamus* samples; the T.S. Smith and Son's Farm (Bridgeville, DE) for providing *A. officinalis* samples. We thank Deepti Ramachandruni and Mayumi Nakano for assistance in data handling and submission, as well as members of the Meyers lab and Virginia Walbot and her lab members (Stanford University) for their helpful discussions. We would like to thank Brewster (Bruce) Kingham, Olga Shevchenko, and Summer Thompson in the University of Delaware Sequencing and Genotyping Facility for their assistance with sequencing.

Funding

This work was supported by funding provided by the U.S. National Science Foundation awards #1339229 to B.C.M and #1611853 to A.E.H, and resources provided by the Donald Danforth Plant Science Center.

Conflict of interest statement. None declared.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. J Mol Biol **215**: 403–410

Axtell MJ (2013). Classification and comparison of small RNAs from plants. Annu Rev Plant Biol **64**: 137–159

Axtell MJ, Meyers BC (2018) Revisiting criteria for plant miRNA annotation in the era of big data. Plant Cell **30**: 272–284

Blevins T, Rajeswaran R, Shivaprasad P V, Beknazariants D, Si-Ammour A, Park HS, Vazquez F, Robertson D, Meins F, Hohn T, et al. (2006) Four plant Dicers mediate viral small RNA biogenesis and DNA virus induced silencing. Nucleic Acids Res **34**: 6233–6246

Cuperus JT, Carbonell A, Fahlgren N, Garcia-Ruiz H, Burke RT, Takeda A, Sullivan CM, Gilbert SD, Montgomery TA, Carrington JC (2010) Unique functionality of 22-nt miRNAs in triggering RDR6-dependent siRNA biogenesis from target transcripts in Arabidopsis. Nat Struct Mol Biol **17**: 997–1003

Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol **16**: 157

Fan Y, Yang J, Mathioni S, Yu J, Yang X, Wang L, Zhang Q, Shen J, Cai Z, Xu C, et al. (2016) *PMS1T*, producing phased small-interfering RNAs, regulates photoperiod-sensitive male sterility in rice. Proc Natl Acad Sci USA **113**: 15144–15149

Fei Q, Li P, Teng C, Meyers BC (2015) Secondary siRNAs from Medicago NB-LRRs modulated via miRNA-target interactions and their abundances. Plant J **83**: 451–465

Fei Q, Xia R, Meyers BC (2013) Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. Plant Cell **25**: 2400–2415

Harkess A, Zhou J, Xu C, Bowers JE, van der Hulst R, Ayyampalayam S, Mercati F, Riccardi P, McKain MR, Kakrana A, et al. (2017). The asparagus genome sheds light on the origin and evolution of a young y chromosome. Nat Commun **8**: 1279

He D, Wang Q, Wang K, Yang P (2015) Genome-wide dissection of the microRNA expression profile in rice embryo during early stages of seed germination. PLoS One **10**: e0145424

Jeong D-H, Park S, Zhai J, Gurazada SGR, De Paoli E, Meyers BC, Green PJ (2011) Massive analysis of rice small RNAs: mechanistic implications of regulated micrornas and variants for differential target RNA cleavage. Plant Cell **23**: 4185–4207

Johnson C, Kasprzewska A, Tennessen K, Fernandes J, Nan GL, Walbot V, Sundaresan V, Vance V, Bowman LH (2009) Clusters and superclusters of phased small RNAs in the developing inflorescence of rice. Genome Res **19**: 1429–1440

Johnson N, Yeoh J, Coruh C, Axtell M (2016) Improved placement of multi-mapping small RNAs. G3 **6**: 2103–2111

Kakrana A, Mathioni SM, Huang K, Hammond R, Vandivier L, Patel P, Arikit S, Shevchenko O, Harkess AE, Kingham B, et al. (2018) Plant 24-nt reproductive phasiRNAs from intramolecular duplex mRNAs in diverse monocots. Genome Res **28**: 1333–1344

Kakrana A, Hammond R, Patel P, Nakano M, Meyers BC (2014) sPARTA: a parallelized pipeline for integrated analysis of plant miRNA and cleaved mRNA data sets, including new miRNA target-identification software. Nucleic Acids Res **42**: e139

Kakrana A, Li P, Patel P, Hammond R, Anand D, Mathioni S, Meyers B (2017) PHASIS: a computational suite for de novo discovery and characterization of phased, siRNA-generating loci and their miRNA triggers. bioRxiv, doi: 10.1101/158832

Kellogg EA (2001) Evolutionary history of the grasses. Plant Physiol **125**: 1198–1205

Komiya R, Ohyanagi H, Niihama M, Watanabe T, Nakano M, Kurata N, Nonomura K-I (2014) Rice germline-specific Argonaute MEL1 protein binds to phasiRNAs generated from more than 700 lincRNAs. Plant J **78**: 385–397

Kozomara A, Griffiths-Jones S (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res **42**: D68–73

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol **10**: R25

Lee WC, Lu SH, Lu MH, Yang CJ, Wu SH, Chen HM (2015) Asymmetric bulges and mismatches determine 20-nt microRNA formation in plants. RNA Biol **12**: 1054–1066

Leung AKL, Sharp PA (2010) MicroRNA functions in stress responses. Mol Cell **40**: 205–215

Li F, Pignatta D, Bendix C, Brunkard JO, Cohn MM, Tung J, Sun H (2011) MicroRNA regulation of plant innate immune receptors. Proc Natl Acad Sci USA **109**: 1790–1795

Liu PP, Montgomery TA, Fahlgren N, Kasschau KD, Nonogaki H, Carrington JC (2007) Repression of AUXIN RESPONSE FACTOR10 by microRNA160 is critical for seed germination and post-germination stages. Plant J **52**: 133–146

Lu C, Jeong D-H, Kulkarni K, Pillay M, Nobuta K, German R, Thatcher SR, Maher C, Zhang L, Ware D (2008) Genome-wide analysis for discovery of rice microRNAs reveals natural antisense microRNAs (nat-miRNAs). Proc Natl Acad Sci USA **105**: 4951–4956

Mallory AC, Dugas DV, Bartel DP, Bartel B (2004) MicroRNA regulation of NAC-domain targets is required for proper formation and separation of adjacent embryonic, vegetative, and floral organs. Curr Biol **14**: 1035–1046

Mathioni SM, Kakrana A, Meyers BC (2017) Characterization of plant small RNAs by next generation sequencing. Curr Protoc Plant Biol **2**: 39–63

May P, Liao W, Wu Y, Shuai B, McCombie WR, Zhang MQ, Liu QA (2013) The effects of carbon dioxide and temperature on microRNA expression in Arabidopsis development. Nat Commun **4**: 1–11

Md Yusuf NH, Ong WD, Redwan RM, Latip MA, Kumar SV (2015) Discovery of precursor and mature microRNAs and their putative gene targets using high-throughput sequencing in pineapple (Ananas comosus var. comosus). Gene **571**: 71–80

Mi S, Cai T, Hu Y, Chen Y, Hodges E, Ni F, Wu L, Li S, Zhou H, Long C, et al. (2008). Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5′ terminal nucleotide. Cell **133**: 116–127

Mirarab S, Nguyen N, Guo S, Wang L-S, Kim J, Warnow T (2015) PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. J Comput Biol **22**: 377–386

Montes RAC, de Fátima Rosas-Cárdenas F, De Paoli E, Accerbi M, Rymarquis LA, Mahalingam G, Marsch-Martínez N, Meyers BC, Green PJ, de Folter S (2014) Sample sequencing of vascular plants demonstrates widespread conservation and divergence of microRNAs. Nat Commun **5**: 3722

Nobuta K, Lu C, Shrivastava R, Pillay M, De Paoli E, Accerbi M, Arteaga-Vazquez M, Sidorenko L, Jeong DH, Yen Y, et al. (2008) Distinct size distribution of endogenous siRNAs in maize: Evidence from deep sequencing in the mop1-1 mutant. Proc Natl Acad Sci USA **105**: 14958–14963

Ono S, Liu H, Tsuda K, Fukai E, Tanaka K, Sasaki T, Nonomura KI (2018) EAT1 transcription factor, a non-cell-autonomous regulator of pollen production, activates meiotic small RNA biogenesis in rice anther tapetum. PLoS Genet **14**: e1007238

Palatnik JF, Allen E, Wu X, Schommer C, Schwab R, Carrington JC, Weigel D (2003) Control of leaf morphotgenesis by microRNAs. Nature **425**: 257–263

Patel P, Mathioni S, Kakrana A, Shatkay H, Meyers BC (2018) Reproductive phasiRNAs in grasses are compositionally distinct from other classes of small RNAs. New Phytol **220**: 851–864

Patel P, Ramachandruni SD, Kakrana A, Nakano M, Meyers BC (2015) miTRATA: a web-based tool for microRNA truncation and tailing analysis. Bioinformatics **32**: 450–452

Puzey JR, Karger A, Axtell M, Kramer EM (2012) Deep annotation of populus trichocarpa microrna from diverse tissue sets. PLoS One **7**: e33034

Reyes JLChua NH (2007) ABA induction of miR159 controls transcript levels of two MYB factors during Arabidopsis seed germination. Plant J **49**: 592–606

Song X, Li P, Zhai J, Zhou M, Ma L, Liu B, Jeong DH, Nakano M, Cao S, Liu C, et al. (2012) Roles of DCL4 and DCL3b in rice phased small RNA biogenesis. Plant J **69**: 462–474

Stamatakis A (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**: 1312–1313

Tamiru M, Natsume S, Takagi H, White B, Yaegashi H, Shimizu M, Yoshida K, Uemura A, Oikawa K, Abe A, et al. (2017) Genome sequencing of the staple food crop white Guinea yam enables the development of a molecular marker for sex determination. BMC Biol **15**: 86

Teng C, Zhang H, Hammond R, Kuang H, Meyers BC, Walbot V (2020) *Dicer-like 5* deficiency confers temperature-sensitive male sterility in maize. Nature Comm **11**: 2912

Xia R, Chen C, Pokhrel S, Ma W, Huang K, Patel P, Wang F, Xu J, Liu Z, Li J, et al. (2019) 24-nt reproductive phasiRNAs are broadly present in angiosperms. Nat Commun **10**: 627

Xia R, Xu J, Arikit S, Meyers BC (2015) Extensive families of miRNAs and *PHAS* loci in Norway spruce demonstrate the origins of complex phasiRNA networks in seed plants. Mol Biol Evol **32**: 2905–2918

Xia R, Xu J, Meyers BC (2017) The emergence, evolution, and diversification of the miR390-TAS3-ARF pathway in land plants. Plant Cell **29**: 1232–1247

Yang Z, Ebright YW, Yu B, Chen X (2006) HEN1 recognizes 21-24 nt small RNA duplexes and deposits a methyl group onto the 2′ OH of the 3′ terminal nucleotide. Nucleic Acids Res **34**: 667–675

You C, Cui J, Wang H, Qi X, Kuo L-Y, Ma H, Gao L, Mo B, Chen X (2017) Conservation and divergence of small RNA pathways and microRNAs in land plants. Genome Biol **18**: 158

Yu G, Smith DK, Zhu H, Guan Y, Lam TTY (2017) GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol **8**: 28–36

Zhai J, Jeong D-H, De Paoli E, Park S, Rosen BD, Li Y, González AJ, Yan Z, Kitto SL, Grusak MA (2011) MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. Nat Lib Med **25**: 2540–2553

Zhai J, Arikit S, Simon SA, Kingham BF, Meyers BC (2014) Rapid construction of parallel analysis of RNA end (PARE) libraries for Illumina sequencing. Methods **67**: 84–90

Zhai J, Zhang H, Arikit S, Huang K, Nan G-L, Walbot V, Meyers BC (2015) Spatiotemporally dynamic, cell-type-dependent premeiotic and meiotic phasiRNAs in maize anthers. Proc Natl Acad Sci USA **112**: 3146–3151

Zhang L, Chia JM, Kumari S, Stein JC, Liu Z, Narechania A, Maher CA, Guill K, McMullen MD, Ware D (2009) A genome-wide characterization of microRNA genes in maize. PLoS Genet **5**: e1000716

Zhang Y, Xia R, Kuang H, Meyers BC (2016) The diversification of plant NBS-LRR defense genes directs the evolution of microRNAs that target them. Mol Biol Evol **33**: 2692–2705